Quantis Empíricos

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

(a) são afetados, de forma exagerada, por valores extremos;

(b) apenas com estes dois valores não temos ideia da simetria ou assimetria da distribuição dos dados.

Para contornar esses fatos, outras medidas precisam ser consideradas.

Indicamos, abaixo, alguns quantis e seus nomes particulares.

$q(0,25) = q_1$: 1º Quartil = 25º Percentil
$q(0,50) = q_2$: Mediana = 2º Quartil = 50º Percentil
$q(0,75) = q_3$: 3º Quartil = 75º Percentil
$q(0,40)$: 4º Decil
$q(0,95)$: 95º Percentil

Exemplo 3.5 Suponha que tenhamos os seguintes valores de uma variável X:

$$15, 5, 3, 8, 10, 2, 7, 11, 12.$$

Ordenando os valores, obtemos as estatísticas de ordem $x_{(1)} = 2$, $x_{(2)} = 3$,..., $x_{(9)} = 15$, ou seja, teremos

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15.$$

Usando a definição de mediana dada, teremos que md $= q(0{,}5) = q_2 = x_{(5)} = 8$. Suponha que queiramos calcular os dois outros quartis, q_1 e q_3. A ideia é dividir os dados em quatro partes:

$$2 \quad \mathbf{3} \quad \mathbf{5} \quad 7 \quad \mathbf{8} \quad 10 \quad \mathbf{11} \quad \mathbf{12} \quad 15$$

Uma possibilidade razoável é, então, considerar a mediana dos primeiros quatro valores para obter q_1, ou seja,

$$q_1 = \frac{3+5}{2} = 4,$$

e a mediana dos últimos quatro valores para obter q_3, ou seja,

$$q_3 = \frac{11+12}{2} = 11{,}5.$$

Obtemos, então, a sequência

$$2 \quad 3 \quad (\mathbf{4}) \quad 5 \quad 7 \quad (\mathbf{8}) \quad 10 \quad 11 \quad (\mathbf{11{,}5}) \quad 12 \quad 15$$

Observe que a média dos $n = 9$ valores é $\bar{x} = 8{,}1$, próximo à mediana.

Exemplo 3.5 (continuação). Acrescentemos, agora, o valor 67 à lista de nove valores do Exemplo 3.5, obtendo-se agora os $n = 10$ valores ordenados:

$$2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15 < 67$$

Agora, $\bar{x} = 14$, enquanto que a mediana fica

$$q_2 = \frac{x_{(5)} + x_{(6)}}{2} = 9,$$

que está próxima da mediana dos nove valores originais, mas ambas (8 e 9) relativamente longe de $\bar{x}$. Dizemos que a mediana é *resistente* (ou *robusta*), no sentido que ela não é muito afetada pelo valor discrepante (ou atípico) 67.

Para calcular q_1 e q_3 para este novo conjunto de valores, considere-os assim dispostos:

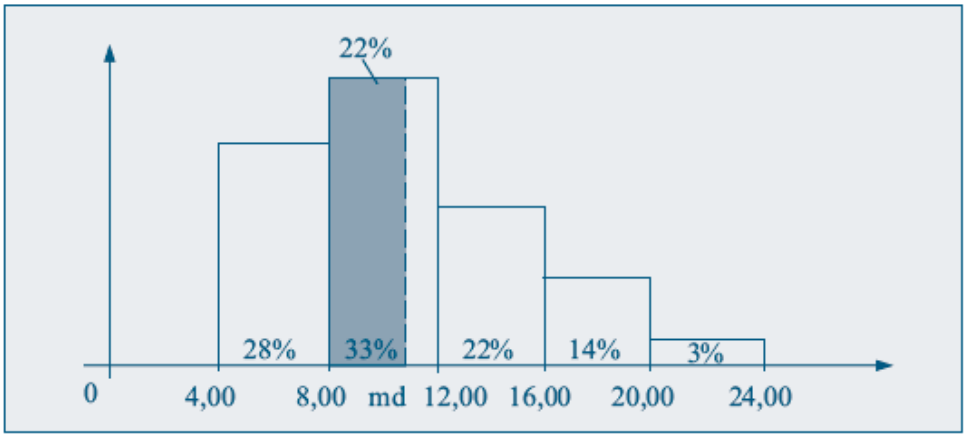
$$2 \quad 3 \quad \mathbf{5} \quad 7 \quad 8 \quad \mathbf{9} \quad 10 \quad 11 \quad \mathbf{12} \quad 15 \quad 67$$

de modo que $q_1 = 5$ e $q_3 = 12$.

Obtemos, então os dados separados em 4 partes por q_1, q_2 e q_3:

$$2 \quad 3 \quad (\mathbf{5}) \quad 7 \quad 8 \quad (\mathbf{9}) \quad 10 \quad 11 \quad (\mathbf{12}) \quad 15 \quad 67$$

Exemplo 3.6 Vamos repetir abaixo a Figura 2.7, que é o histograma da variável S = salário dos empregados da Companhia MB.



$$\frac{12,00 - 8,00}{33\%} = \frac{\text{md} - 8,00}{22\%}$$

ou

$$\text{md} - 8,00 = \frac{22\%}{33\%} \cdot 4,00,$$

logo

$$\text{md} = 8,00 + 2,67 = 10,67,$$

que é uma expressão mais precisa para a mediana do que a mediana bruta encontrada anteriormente.

(a) q(0,25): Verificamos que q(0,25) deve estar na primeira classe, pois a proporção no primeiro retângulo é 0,28. Logo,

$$\frac{q\left(0,25\right)-4,00}{25\%}=\frac{8,00-4,00}{28\%},$$

e então

$$q\left(0,25\right)=4,00+\frac{25}{28}4,00=7,57.$$

(b) q(0,95): Analisando a soma acumulada das proporções, verificamos que este quantil deve pertencer à quarta classe, e que nesse retângulo devemos achar a parte correspondente a 12%, pois a soma acumulada até a classe anterior é 83%, faltando 12% para atingirmos os 95%. Portanto,

$$\frac{q\left(0,95\right)-16,00}{12\%}=\frac{20,00-16,00}{14\%},$$

logo

$$q\left(0,95\right)=16,00+\frac{12}{14}\times 4=19,43.$$

(c) q(0,75): De modo análogo, concluímos que o terceiro quantil deve pertencer ao intervalo 12,00 ⊣ 16,00, portanto

$$\frac{q\left(0,75\right)-12,00}{14\%}=\frac{16,00-12,00}{22\%}$$

e

$$q\left(0,75\right)=14,55.$$

Uma medida de dispersão alternativa ao desvio padrão é a ***distância interquartil***, definida como a diferença entre o terceiro e primeiro quartis, ou seja,

$$d_q = q_3 - q_1. \qquad (3.13)$$

Para o Exemplo 3.5, temos $q_1 = 4$, $q_3 = 11{,}5$, de modo que $d_q = 7{,}5$. Para um cálculo mais preciso, veja o Problema 17. Lá obtemos $q_1 = 4{,}5$, $q_3 = 11{,}25$, logo $d_q = 6{,}75$.

Os quartis $q(0{,}25) = q_1$, $q(0{,}5) = q_2$ e $q(0{,}75) = q_3$ são **medidas de localização resistentes** de uma distribuição.

Dizemos que uma medida de **localização ou dispersão é resistente** quando for pouco afetada por mudanças de uma pequena porção dos dados.

A mediana é uma medida resistente, ao passo que a média não o é.

Os cinco valores, $x_{(1)}$, q_1, q_2, q_3 e $x_{(n)}$, são importantes para se ter uma boa ideia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, deveríamos ter:

(a) $q_2 - x_{(1)} \simeq x_{(n)} - q_2$;

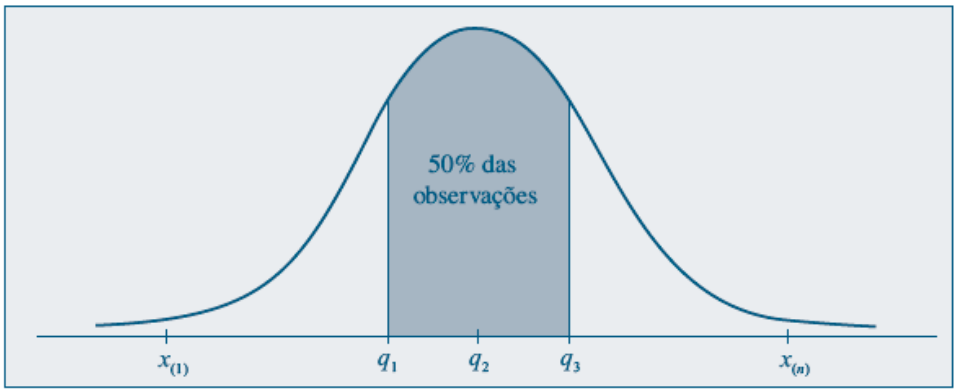
(b) $q_2 - q_1 \simeq q_3 - q_2$;

(c) $q_1 - x_{(1)} \simeq x_{(n)} - q_3$;

(d) distâncias entre mediana e q_1, q_3 menores do que distâncias entre os extremos e q_1, q_3.

A diferença $q_2 - x_{(1)}$ é chamada *dispersão inferior* e $x_{(n)} - q_2$ é a *dispersão superior*. A condição (a) nos diz que as duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica.

Figura 3.1 Uma distribuição simétrica: normal ou gaussiana.



Na Figura 3.2, temos ilustradas estas cinco medidas para os $n = 9$ valores do Exemplo 3.5.

Figura 3.2 Quantis e distâncias para o Exemplo 3.5.

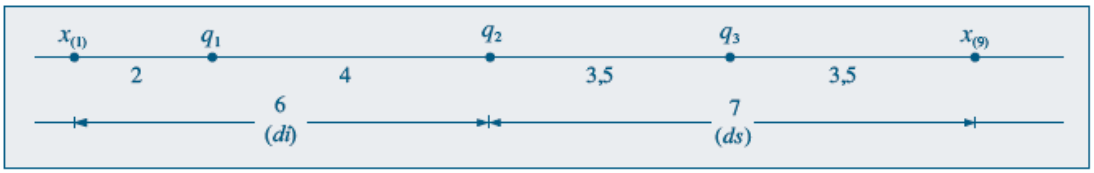
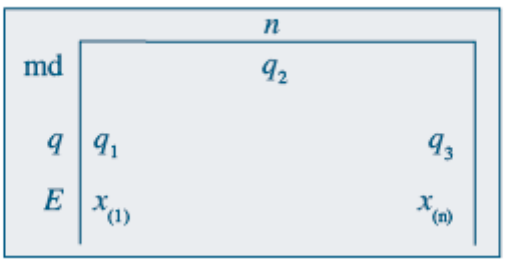
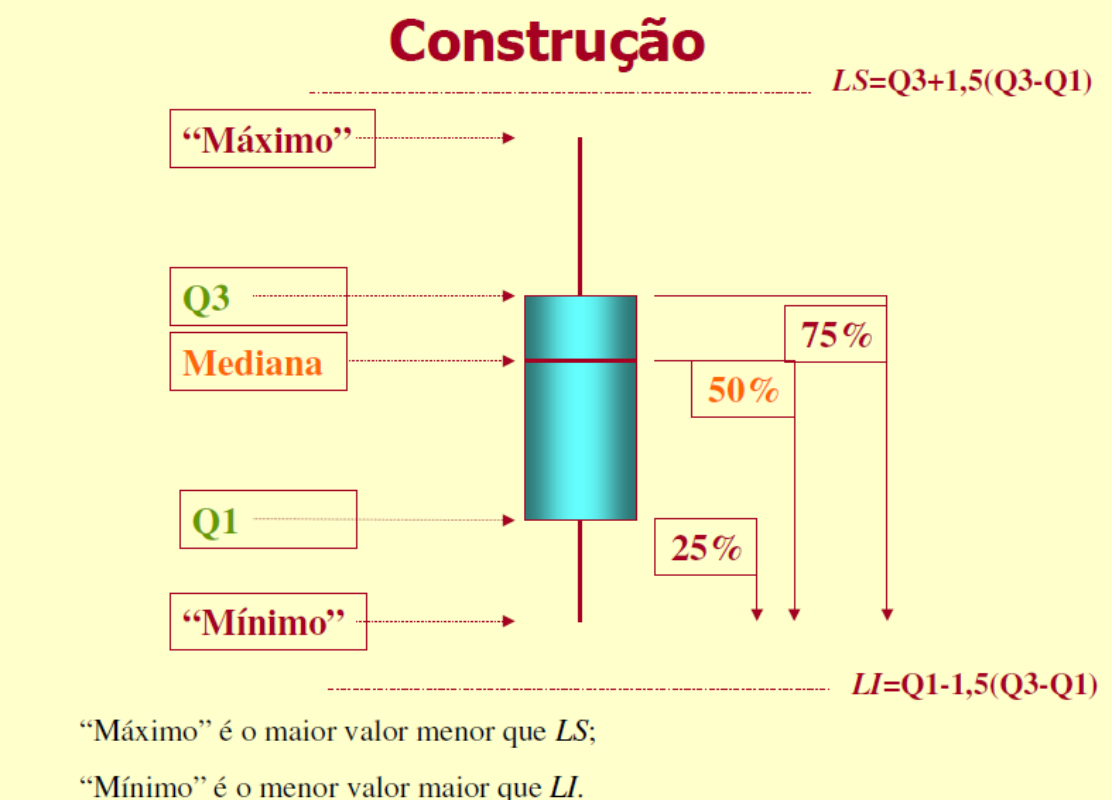


Figura 3.3 Esquema dos cinco números.



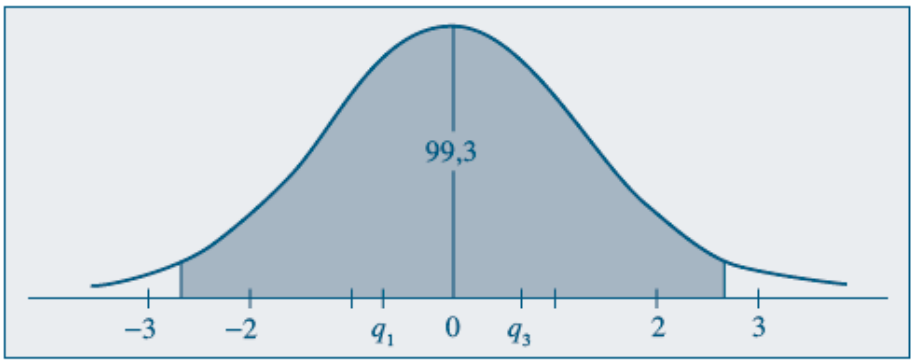
Box Plots



"Máximo" é o maior valor menor que *LS*;

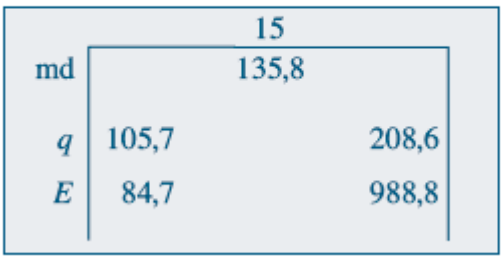
"Mínimo" é o menor valor maior que *LI*.

Figura 3.8 Área sob a curva normal entre LI e LS.



Exemplo 3.8 15 maiores municípios do Brasil, ordenados pelas populações.

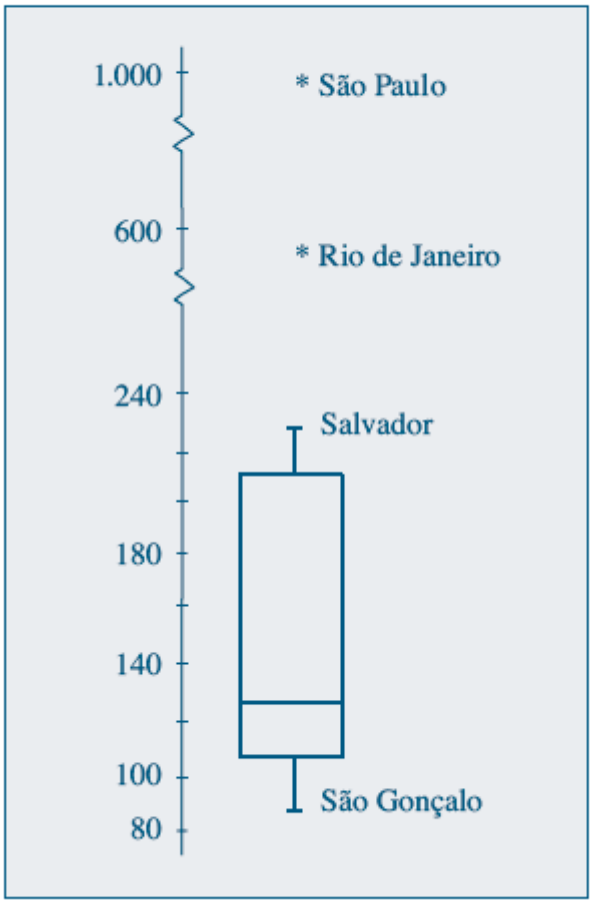
Figura 3.5 Esquema dos cinco números para o Exemplo 3.8.



$$\text{LI} = q_1 - (1{,}5)d_q = 105{,}7 - (1{,}5)\ (102{,}9) = -\ 48{,}7,$$

$$\text{LS} = q_3 + (1{,}5)d_q = 208{,}6 + (1{,}5)\ (102{,}9) = 362{,}9.$$

Figura 3.6 *Box plot* para os quinze maiores municípios do Brasil.

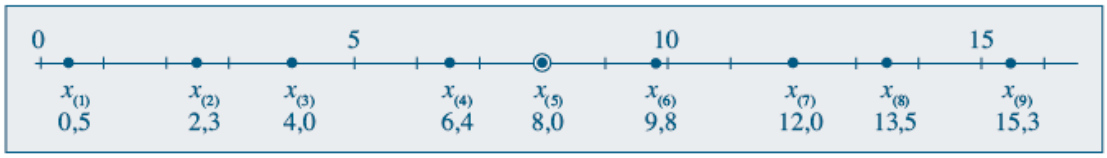


Gráficos de Simetria

Figura 3.9 Distribuições assimétricas.

Exemplo 3.9 Considere os dados que, dispostos em ordem crescente, ficam representados no eixo real como na Figura 3.10.

Figura 3.10 Dados aproximadamente simétricos.



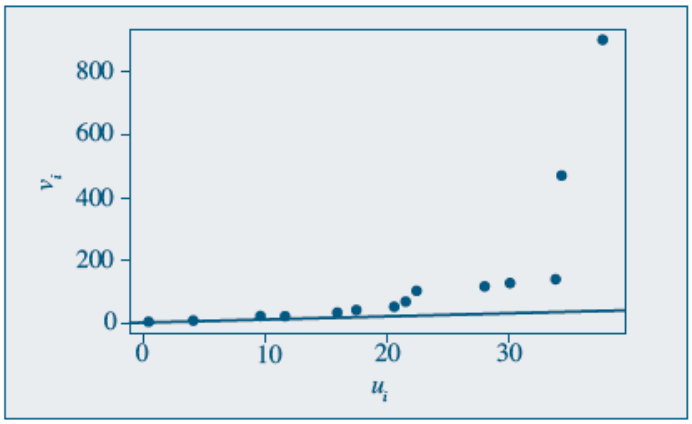
Esses dados são aproximadamente simétricos, pois como $q_2 = 8$, $u_i = q_2 - x_{(i)}$, $v_i = x_{(n+1-i)} - q_2$, teremos:

$$u_1 = 8{,}0 - 0{,}5 = 7{,}5, \quad v_1 = 15{,}3 - 8{,}0 = 7{,}3,$$
$$u_2 = 8{,}0 - 2{,}3 = 5{,}7, \quad v_2 = 13{,}5 - 8{,}0 = 5{,}5,$$

$$u_3 = 8{,}0 - 4{,}0 = 4{,}0, \quad v_3 = 12{,}0 - 8{,}0 = 4{,}0,$$
$$u_4 = 8{,}0 - 6{,}4 = 1{,}6, \quad v_4 = 9{,}8 - 8{,}0 = 1{,}8.$$

A Figura 3.11 mostra o gráfico de simetria para as populações dos trinta municípios do Brasil. Vemos que a maioria dos pontos estão acima da reta $v = u$, mostrando a assimetria à direita da distribuição dos valores. Nessa figura, vemos destacados os pontos correspondentes a Rio de Janeiro e São Paulo.

Figura 3.11 Gráfico de simetria para o CD-Municípios.



Transformações

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \ell n(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0. \end{cases} \qquad (3.15)$$

Normalmente, o que se faz é experimentar valores de p na sequência

$$\ldots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \ldots$$

e para cada valor de p obtemos gráficos apropriados (histogramas, *box plots* etc.) para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p.

Exemplo 3.10 Consideremos os dados das populações do CD-Municípios e tomemos alguns valores de p: 0, 1/4, 1/3, 1/2. Na Figura 3.12, temos os histogramas para os dados transformados e, na Figura 3.13, os respectivos *box plots*. Vemos que $p = 0$ (transformação logarítmica) e $p = 1/3$ (transformação raiz cúbica) fornecem distribuições mais próximas de uma distribuição simétrica.

Figura 3.12 Histogramas para os dados transformados. CD-Municípios.

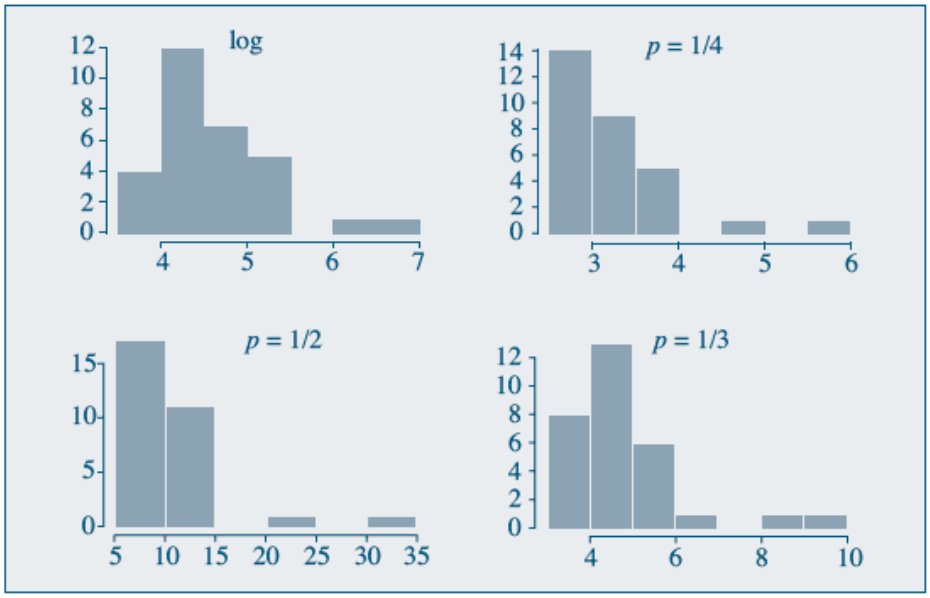


Figura 3.13 *Box plots* para os dados transformados. CD-Municípios. SPlus.